Translating the XNLI RTE pairs to the Moroccan Arabic Dialect (Darija) using a LLM

Hiba El Oirghi
University of Maryland
eloirghi@umd.edu

Abstract

Hundreds of millions in the Middle East and North Africa (MENA) and worldwide write and speak using one of the many variations of Arabic known as dialects (Gregory et al., 2021). These dialects lack standardized spelling and are not recognized as official languages in any of the Arabic-speaking countries in the MENA region (Habash et al., 2005; Gregory et al., 2021). To further develop robust and useful Natural Language Processing (NLP) tools and language models for these underserved populations, I turned to Natural Language Inference (NLI) and augmented the XNLI (Conneau et al., 2018) dataset with the Moroccan Arabic Dialect (Darija) using a large language model - Llama2-7B (Touvron et al., 2023) to perform Machine Translation (MT). Further inspection of the MT output reveals linguistic weaknesses and a large gap in the model's fluency in Darija which is spoken by roughly 40 million people in Morocco (Gregory et al., 2021). Additionally, I also evaluated the model's NLI capabilities on XNLI pairs and manually translated 30 XNLI pairs into Darija to provide gold standards for future research. I hope that this limited work will inspire further research into the fluency of several state-of-the-art models in Arabic dialects and other widely spoken languages and dialects around the world.

1 Introduction

Recognizing textual entailment (RTE) is a powerful task to test for NLI capabilities, with some researchers arguing for the necessity of training and evaluating on RTE pairs in multiple languages to advance language models' capabilities for common-sense reasoning cross-lingually (Conneau et al., 2018).

Below is a summary of the contributions of this work:

- The evaluation of Llama2-7B on English and Modern Standard Arabic (MSA) XNLI pairs;
- The augmentation of the XNLI dataset with Moroccan Arabic Dialect (Darija) translations using Llama2-7B;
- The manual translation of 30 XNLI pairs from English and MSA to Darija;
- The analysis of Llama2-7B's performance in translating between high-resource (English, MSA) and low-resource (Darija) languages.

2 RTE Performance on the XNLI Dataset

The XNLI Dataset

The XNLI dataset is a subset of 5k-ish from Multi-Genre Natural Language Inference (MNLI) translated into 14 different languages including some low-ish resource languages such as MSA and Urdu (Conneau et al., 2018). The XNLI dataset is mainly used for the training and evaluation of a NLI classification task: the prediction of textual entailment label: does the premise imply, contradict, or neither the hypothesis? The task is simplified as follows: given two sentences, predict one of three labels. The XNLI dataset subset I use in this paper is structured in rows such as:

- Premise: a multilingual string variable in English and Arabic;
- Hypothesis: a multilingual string variable in English and Arabic;
- Label: a classification label, with possible values 0: entailment, 1: neutral, 2: contradiction

Performance of Llama2-7B on RTE task

I evaluated Llama2-7b on the English and MSA subsets of the XNLI dataset. The model demonstrated weak performance - **39.98%** and **33.33%** accuracy for English and Arabic pairs, respectively, mainly due to its overall inability to respond with a one-word label despite being prompted to do so as

follows:

Given the following premise and hypothesis, determine if the premise entails the hypothesis, contradicts it, or neither (neutral). Respond with only one of these labels: entailment, contradiction, or neutral.
Premise: premise in English or MSA
Premise: hypothesis in English or MSA
Label:

The slightly higher performance on English pairs can be explained by the reasonable assumption that the model was trained with larger volumes of English data compared to Arabic data. This gap in performance speaks to the fluency gap LLMs face when dealing with languages other than English, even a medium-resource language such as MSA in this case.

The overall low performance on English and Arabic XNLI pairs for the RTE task can be explained by the model's inability to output one single label in response to the evaluation prompt, despite my experimentation with various prompts and model settings. In future work, I intend to use state-of-the-art models to test their performance on RTE tasks.

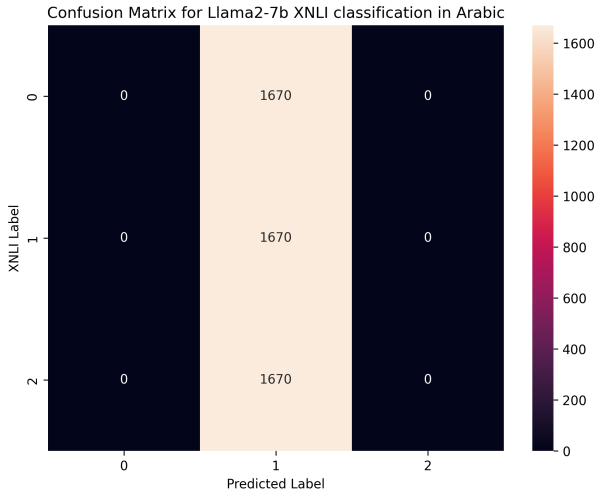

Figure 1: Confusion matrix for Arabic XNLI pairs using LLaMA2 7B for label prediction. The labels are: "entailment": 0, "neutral": 1, "contradiction": 2.

3 Data Collection

Human Translation of a subset of XNLI

I followed the following rules when translating 30 XNLI pairs (60 sentences) from English and Arabic to Darija. First, I asked "How would a native speaker of Moroccan Arabic (Darija) say this sentence and how would they spell it?", this approach is inspired by the creation of the MADAR parallel corpus (Bouamor et al., 2018) where they focused on authenticity when annotating they parallel corpus for Arabic dialects. Additionally, previous work analyzing human translation from MSA to Arabic dialects (Obeid et al., 2018) has shown that native Arabic dialects annotators are more likely to use inauthentic language when translating from MSA to an Arabic dialect - this is why I opted for English as a source language and only consulted the MSA XNLI pairs to ensure that I was using similar language when translating foreign entities (e.g AFFC Air Force, Augusta, GA). Second, and unlike the approach in previous prominent works (Bouamor et al., 2018), I did not follow a specific set of orthographic rules and simply spelled Darija the way I would in a daily normal setting. This approach has gained popularity recently (Talafha et al., 2024) as a way to preserve the authenticity of dialectical spelling and to avoid wrongfully instituting spelling rules for low-resource 'speech communities' (Bird and Yibarbuk, 2024). Except for terms that were kept in Latin script (e.g. AFFC, U2) in the original MSA XNLI subset, I did not change the code to French - a common strategy among native Darija speakers (Talafha et al., 2024) - and used the singular masculine form in case of ambiguity. The manual translation took me approximately 3 hours to finish.

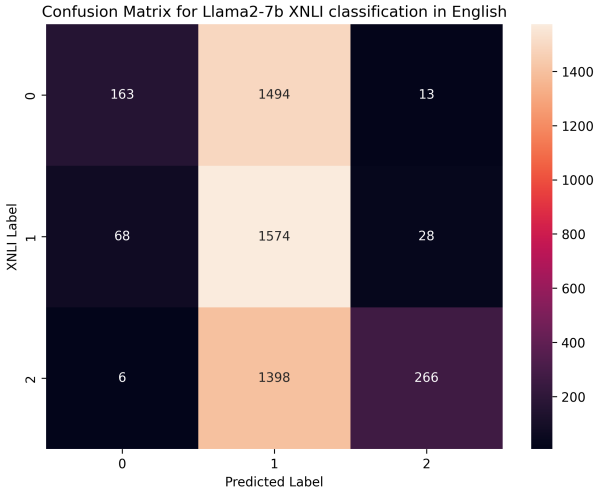

Figure 2: Confusion matrix for English XNLI pairs using LLaMA2 7B for label prediction. The labels are: "entailment": 0, "neutral": 1, "contradiction": 2.

XNLI Augmentation using LLMs for MT

To augment 5010 XNLI pairs from the training dataset to Darija, I leveraged Llama2-7b's translation capabilities to perform two MT tasks:

- English to Darija
- MSA to Darija

For each MT language pair, I prompted Llama2-7b to translate the premise and hypothesis separately, maintaining the original label. I used one-shot prompting and experimented with several prompts to improve the quality of MT output in Darija. The MT prompt is as follows:

Translate the following Source Language: MSA or English sentence to Moroccan Arabic. Only respond with the translated sentence in Arabic letters.
Arabic: Premise or Hypothesis in Source Language
Moroccan Arabic:

Upon manual inspection of the MT output by Llama2-7B, I discovered that the main issues in the MT output include the use of Latin Script, the regurgitation of the prompt, the use of emojis, notes about the colloquial nature of Darija, and defaulting to MSA and English terms. Additionally, the model consistently produced translation output that mixed Darija with French or English words, despite the prompt specifically asking for only output in the Arabic script.

To broadly examine the MT output, I cleaned up the raw output from Lama2-7B by removing all latin script and redundant punctuation. The average token length for the MT output in the case of Arabic and English as source languages is comparable to that of the average taken length for the humanly produced English and Arabic sentences in the XNLI dataset (See Table 1), of course, this does not guarantee the model's fluency or the quality of the MT.

	XNLI English	XNLI MSA	MT English - Darija	MT MSA - Darija	HT English - Darija
Premise	21.7	20.7	18.5	32.3	13.2
Hypothesis	10.7	10.2	9.2	16.9	7.2

Table 1: Average number of tokens per sentence. (Conneau et al., 2018)

Limitations

This work is an exploratory probe into data augmentation through LLM-based machine translation. The main limitations can be classified into three categories: limited resources, limited expertise, and scope of the question. First, the work was constrained due to limited time, access to native Arabic dialect speakers, and limited computing resources. Translating 5010 XNLI pairs took about 8 hours on the Nexus GPU cluster. As the author and only human translator, I gained valuable insight into the human translation process but the resulting translated pairs most likely suffer from well-defined problems such as dataset artifacts (Poliak et al., 2018; Geva et al., 2019). Second, due to the solo nature of this work, it naturally suffers from a limited scope and can be broadened to include a more in-depth error analysis. This work is not intended to make any claims about the model (Touvron et al., 2023) used for NLI evaluation or MT but rather to expose the gap in performance between dominant high-resource languages (e.g. English) and unstandardized low-resource languages such as Darija.

Ethics Statement

This work is intended as an exploration of a LLM's abilities for MT to Darija as well as its ability to perform a RTE task. I do not intend to imply that MT using LLMs should replace human translation. I hope that future NLP research for low-resource languages is inclusive with a central focus and respect for 'Speech Communities' at the expense of blind generalizability (Bird and Yibarbuk, 2024).

Future Work

Plans for future work can be classified into three categories to address the abovementioned limitations. To address limited resources and expertise, I would like to collaborate with other researchers to build on this paper by expanding the LLMs used for evaluation to the Llama3 Instruct models (Grattafiori et al., 2024), LLMs fine-tuned specifically for machine translation tasks such as TowerInstruct-7B (Alves et al., 2024), as well as specifically MSA-English LLMs such as Jais-7B (Sengupta et al., 2023). Additionally, I plan on comparing the MT translation outputs with human translation. This work will be executed over months and benefit from our combination of skills, compute power, and extended man-hours. In a more exploratory vein, I would like to experiment with fine-tuning the best-performing model on the Darija subset of the Casablanca dataset - the largest fully segmented and annotated Arabic dialect dataset

with an impressive 48 hours of transcribed audio covering 8 MENA dialects (Talafha et al., 2024). Future research will also aim to explore better definitions for Arabic dialects (i.e. what constitutes a dialect?) and benefit from more human translation of existing NLI datasets into various Arabic dialects.

Acknowledgements

This work was inspired by the enriching conversations in class - CMSC 828J. The instructing team and the students created an inclusive environment where every session lead to productive discussion. Thank you for a great semester.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias

Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Laura Gregory, Hanada Taha Thomure, Amira Kazem, Anna Boni, Mahmoud Elsayed, and Nadia Taibah. 2021. *Advancing Arabic Language Teaching and Learning: A Path to Reducing Learning Poverty in the Middle East and North Africa*.

Nizar Habash, Owen Rambow, and George Kiraz. 2005. Morphological analysis and generation for Arabic dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24, Ann Arbor, Michigan. Association for Computational Linguistics.

Ossama Obeid, Salam Khalifa, Nizar Habash, Houda Bouamor, Wajdi Zaghouani, and Kemal Oflazer.

2018. MADARi: A web interface for joint Arabic morphological annotation and spelling correction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, Hour Mohamed, Fakhraddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, Sara Shatnawi, Yasir Ech-chammakhy, Amal Makouar, Yousra Berrachedi, Mustafa Jarrar, Shady Shehata, Ismail Berrada, and Muhammad Abdul-Mageed. 2024. Casablanca: Data and models for multidialectal Arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.